Predicting Wages and Recruitment Rates in Swedish Regions and Municipalities

Youheng Lü
Department Microdata Analysis
Dalarna University
Borlänge, Sweden
h21youlu@du.se

***Abstract*—Finding the right place to look for a well-paying job can take a lot of time that would be better suited elsewhere. This paper tackles this problem by creating models for the predicting wages and recruitment rates in Swedish Regions and Municipalities. It establishes a Linear Regression prediction model. Using 5-fold Cross-Validation to compare the accuracy against Tree-based models, this paper provides a baseline model that can be improved on for future work. The key takeaways are that a Linear Regression model is accurate at predicting Wages in Sweden with a CV-Score of 0.89. The other takeaway is that a Linear Regression model for predicting recruitment rates had very low accuracy and is not suited for prediction in it's current form.**

***Index Terms*—linear regression, wage prediction, recruitment rate, decision tree, random forest**

I. Introduction

Gathering the necessary information to find out where to look for a job is a lot of work. This question is especially important for future University graduates are looking for opportunities to find well-paying employment in Sweden. To do that a model which can predict wages and recruitment rates based on region and municipality should raise the chances of finding a well-paying job relatively quickly.

On the topic of predicting wages work has been done using techniques for imputing missing pay data and forecasting salaries over a 6-year period. [2]. The paper developed numerous model specifications based on the link between salaries and other macroeconomic and labour market factors, and then employed pseudo out-of-sample testing to pick the top performing models by nation. This technique allowed for the production of a balanced panel data collection for 112 nations throughout the period 1995-2019. Another work focussed specifically on Sweden. The paper did analysis of the relations between private and public sector wages. It discovered two long-run correlations between central government, local government, and private sector salaries in Sweden using a maximum likelihood cointegration technique [3]. Other work predicted the minimum salary with Wavelet analysis and adaption methods in EU-Member states [4].

Work related to employment prediction exists in the following areas: One paper analysed the effectiveness of a model based on job search-related web queries in predicting quarterly unemployment rates in small samples [5]. Work has also been done on the analysis of how Employment is distributed in Sweden over the Years, showing that local government employment rose from 1959 - 1993 [6]. Another paper focussed a probabilistic analysis of the likelihood on finding a job [7] given the length of unemployment. The result show that people with longer unemployment have lower chances of finding employment.

In the end we did not find any work that builds a model that helps jobseekers predict the current environment to find the best opportunity for a job. We therefore use the following: Statistics Sweden [8] provides multiple datasets about the Swedish labour market. Here, we analyse two datasets: The average wage development of the municipalities over the years 2007 - 2020 and the quarterly recruitment rates for different regions in Sweden from 2015 - 2021. The goal is to create a predictive Model to show the development of Wages and recruitment rates in the future. This should help students decide

TABLE I
DESCRIPTION OF AVERAGE MONTHLY SALARY DATASET

Feature	Type	values
sex	categorical	”men”, ”women”, ”total”
region	categorical	292 unique municipalities
year	numerical	2007 - 2020
average wage	numerical	avg. monthly wage in SEK

TABLE II
DESCRIPTION OF RECRUITMENT RATE DATASET

Feature	Type	values
region	categorical	9 unique regions
quarter	string	2015K2 - 2021K4
recruitment rate	numerical	Recr. rate for Business sector

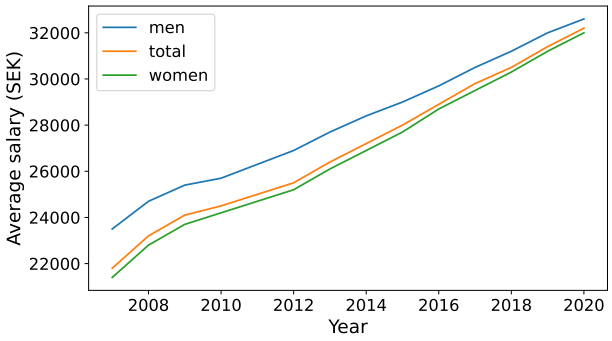

Fig. 1. Average monthly Salary in Sweden, split by men, women and total.

in which region to find employment. We will answer the following two research questions:

A. Research Questions

1) Can we create a model that accurately predicts wages for Sweden’s municipalities?
2) Can we create a model that accurately predicts recruitment rates for the municipality?

We organize the paper as follows. The Methods and Material section provide a general overview of the dataset and the methods used to build and evaluate the prediction model.

II. METHODS AND MATERIAL

For the wage prediction we use the dataset ”Average monthly salary in the municipalities by municipality and sex. Year 2007 - 2021” [1]. This dataset contains 12264 rows and 4 columns. A detailed description can be found in Table I.

For the prediction of recruitment rate, we use the dataset ”Recruitment and vacancy rate, Business by region NUTS2, Quarter 2015K2 - 2021K4” [2]. The dataset also contains observations for vacancy rate but since we are only interested in recruitment rate, we drop them. The dataset then contains 243 rows and 3 columns. A detailed description can be found in Table II.

A. Model

Andreas Jokobsson has provided a general overview of Time-Series modelling [1]. His book describes Multiple methods how to model time-Series data. This paper uses his methodology described in Chapter 2.2.4 on Linear Projections on normal distributed vectors to make Predictions of the future. We want to see whether a Linear regression model can fit, since it is easy to scale for time series data, where only the year will change, and all the other variables stay the same. To see whether that applies we first plot the data to gain first insights. After gaining those insights, we modify the data accordingly for the model to be applicable. On this data we create a Linear Regression model. To evaluate the model, we use 5-fold Cross Validation and use the CV-Score metric from the *skikit-learn* library. We compare the CV-Scores of Linear Regression to a Decision Tree and a Random-Forest with Hyper parameter Tuning. The CV-Score of the Decision tree provides a baseline how good the model can be since it does not have any assumptions about the distribution of the data. The Random-Forest should perform better than the Decision tree, so we also fit that. We only use the tree based models for evaluation since we expect them to perform badly for predicting Time-Series data. Given the output is based on decision nodes, all data of the future will fall through the same node (e.g years >2020) which means that there are no changes in the prediction.

[1]https://www.statistikdatabasen.scb.se/goto/en/ssd/Kommun17g
[2]https://www.statistikdatabasen.scb.se/goto/en/ssd/KV15LJVAKgrreg

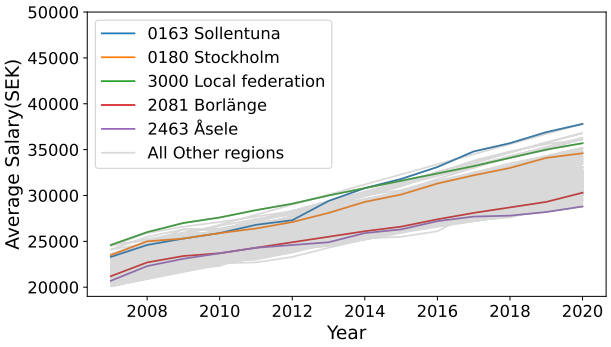

Fig. 2. Average monthly Salary in selected regions. Sollentuna and Hällefors start at a similar wage in 2007 but reach quite different endpoints, Laxa has highest wage in 2007 but is low in 2020. We choose Stockholm and Borlänge since the audience should be familiar with the two municipalities. The grey lines show the wages for all other regions.

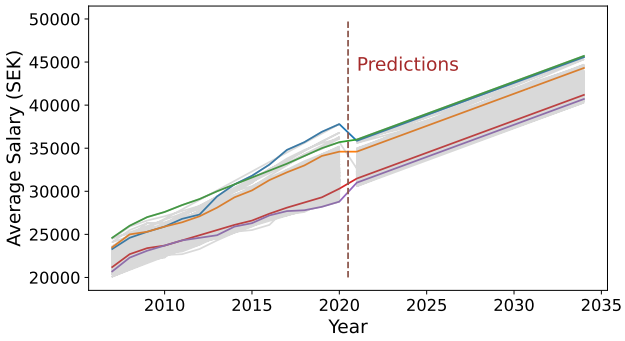

Fig. 3. Prediction of **total** Wages for all the municipalities from 2021 - 2035. The same regions like in Figure 2 have been chosen, so the legend was removed to prevent distraction from the data.

TABLE III
MODEL COMPARISON WAGE PREDICTION

Model	**CV-Score**	**Variance**
Linear Regression	0.91	0.04
Decision Tree	0.91	0.06
Random Forest	0.91	0.02
Tuned Random Forest	0.91	0.02

TABLE IV
TOP 5 MUNICIPALITIES WITH THE HIGHES WAGES IN 2027

Region	**Year**	**Average monthly salary**
3000 Local federation	2027	40475.922400
0160 Täby	2027	40368.779542
0163 Sollentuna	2027	40333.065257
0182 Nacka	2027	40083.065257
0114 Upplands Väsby	2027	39461.636685

III. RESULTS

A. Can we create a model that accurately predicts wages for Sweden's municipalities?

Our goal here was to build an accurate model to predict salaries in Sweden for next few years. To do that we first plot the data to find interesting patterns that can be used for modelling.

A first look at the average wage in all of Sweden Fig. 1 shows that wages have been increasing at a very steady rate. We can also see that the wage gap between men and women grows closer between the years. For sake of simplicity, since the data could affect our Cross-Validation, we will drop the *sex* column and only look at the total wages.

In Figure 2 we can see that there is a general linear upwards trend in the municipalities, which shows that Linear Regression model should fit the data well. On the other hand, we can also see that different municipalities can have different trajectories in the case of Sollentuna versus Hällefors, where they both start at a similar wage but reach a 10000 SEK difference in 2020.

We applied One-Hot-encoding on the column *region* so our model can work with this categorical data. In TABLE III Linear Regression has a score of 0.89 out of a maximum of 1. It has comparable results to a Decision Tree and Random Forest model which suggest that the model is already accurate and suitable for prediction.

The result of the Linear Regression model for wage prediction is shown in Figure 3. As we can see the model predicts based not only on the average wage from 2020 but from the whole data range from 2007 to 2020 of the municipality. The strength of the approach is that it takes the historical development into account and will predict accurately in the future. One weakness could be that the values for 2021, 2022 might be inaccurate, which can be seen in case of Sollentuna (blue), where the prediction shows a jump in Wages in 2021 for some regions that stems from taking the average over the years as the predictor. To give general overview how the predictions look like we show the top 5 regions with the highest wage in 2027 in TABLE IV.

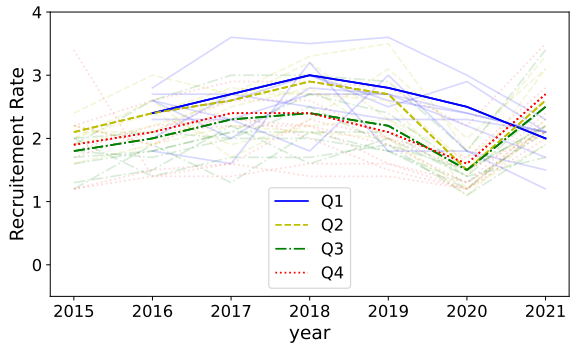

Fig. 4. Recruitment rate in the 9 unique regions of Sweden by quarter. The strong line represents all of Sweden while the lines in the background are the quarterly recruitment rates for all the regions.

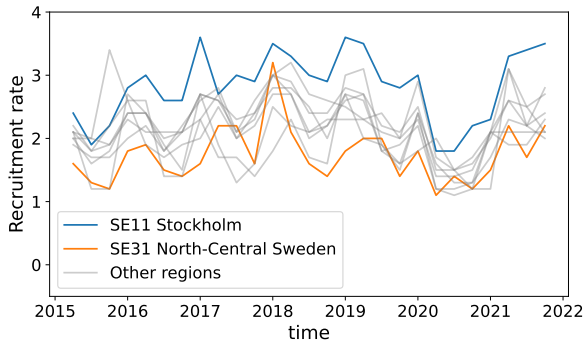

Fig. 5. Recruitment rate in the 9 unique regions of Sweden by total time. The Regions SE11 Stockholm and SE31 North-Central Sweden have been highlighted since they should be most familiar to the target audience of this paper.

B. Can we create a model that accurately predicts recruitment rates for the municipality?

We first split the column *Quarter* into two columns *Year* and *Quarter* to be able to work on those numerically. We then create a new variable called *Time* that has the formula:

$$Time = Year + (Quarter - 1) \cdot 0.25 \quad (1)$$

which translates each quarter as the beginning of the quarter of the year for simplicity. We propose two linear models to predict the response variable Y (*Recruitment rate*) can take two forms:

$$Y = a_{Intercept} + a_1 \cdot Year + a_{quarter} + a_{region} \quad (2)$$

This first model assumes that there are quarterly differences, so we use the quarters as a categorical variable or Form 2:

$$Y = a_{Intercept} + a_1 \cdot time + a_{region} \quad (3)$$

The second model minimizes variance by cutting down on features: I use the variable *time* instead. This model is supposed to perform well if a general trend like in the Wages data exists. It also has the advantage of being the simpler model, which given our small dataset might be a better fit.

In figure 4 we look compare the different Quarters. We can see that Q1 and Q2 have higher recruitment rate than Q3 and Q4. We see the fall in recruitment rate in 2020 during the COVID-19 Pandemic as well as the subsequent rise in 2021. It is hard to pinpoint specific trends over the quarters, since the data has big variation, especially during Q1. To see if there is a general trend, we plot the development of Recruitment rates over time. In Figure 5 we can see that there are times where most regions recruit more and times where most regions recruit less. We can also see a clear difference between the region SE11 Stockholm and SE31 North-Central Sweden. This suggests that region vary in the height of the employment rate. We can't find specific pattern that are useful for modelling given that there is also high variance in the data. Adding the fact that our dataset is also small, we can expect models to perform poorly.

TABLE V
COMPARISON OF RECRUITMENT RATE PREDICTION MODEL

Model	**CV-Score**	**Variance**
Linear Regression Equation (2)	-0.03	0.23
Decision Tree Equation (2)	-0.17	0.33
Liner Regression Equation (3)	-0.18	0.20
Decision Tree Equation (3)	0.00	0.24

We fit the models based on Equation (2) and Equation (3). The results are show in TABLE V. As expected, we can see that all models perform badly, with CV-Scores of zero or below and high Variance. Decision-Trees do not perform good either, we therefore conclude that the current data is not enough to make accurate predictions.

IV. CONCLUSION

We observed a general linear upwards trend for wage data with the wage difference between men and women becoming smaller over the years. Our

Linear Regression model performs well with a high CV-Score. It even has similar performance relative to Decision Tree and Random Forest which we used for evaluation. Using this model for prediction, we can see that it captures the general trend of the data well. One caveat is the prediction in the first few years, where there are big jumps between 2020 and 2021. The model should therefore perform best for medium-term predictions around 5 years in the future. To improve this model Moving average smoothing can be applied so predictions for short-term are smoothed and more accurate.

With our recruitment rate dataset, we found that the data shows no clear pattern and has high variance. We tried two approaches: Splitting the data by quarters and modelling just with the *time* feature. Both approaches worked poorly, with low CV-scores and high Variance in the model. The models created are therefore not suited for prediction. Still, we made some general observations that should be helpful: Firstly, Q1 and Q2 have higher recruitment rates than Q3 and Q4. Second SE11 Stockholm has comparatively higher recruitment rates than the rest of Sweden.

To improve on this model for predicting recruitment rates, we propose the following steps: The first step would be to find a more comprehensive dataset. There is a dataset that also includes vacancy rate of the regions which could be added. Another way would be to include a dataset of municipalities that contains more than the nine regions of the current dataset. With a more comprehensive dataset, researchers could consider other more comprehensive modelling approaches that consider the trend and seasonality of the data. Moving average smoothing should also be considered.

REFERENCES

[1] Kitagawa, Genshiro. Introduction to time series modeling. Chapman and Hall/CRC, 2010.

[2] Ernst, Ekkehard, et al. "Predicting Wages." (2016).

[3] Jacobson, T., Ohlsson, H. Long-run relations between private and public sector wages in Sweden. Empirical Economics 19, 343–360 (1994). https://doi.org/10.1007/BF01205942

[4] Hadas-Dyduch, Monika. Model-spatial approach to prediction of minimum wage. No. 29/2016. Institute of Economic Research Working Papers, 2016.

[5] Francesco, D'Amuri. "Predicting unemployment in short samples with internet job search query data." (2009).

[6] Rosen, Sherwin. "Public Employment and the Welfare State in Sweden." Journal of Economic Literature, vol. 34, no. 2, 1996, pp. 729–40. JSTOR, http://www.jstor.org/stable/2729220. Accessed 20 May 2022.

[7] Shimer, Robert. "The probability of finding a job." American Economic Review 98.2 (2008): 268-73.

[8] Statistics Sweden, Weblink: https://www.scb.se/en/